EFFICIENCY PREDICTION AND MECHANISM DISCOVERY FOR THE CRISPR-CAS9 SYSTEM

A Thesis

by

YI YAN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Yang Shen
Committee Members,	Jean-Francois Chamberland-Tremblay
	Jim Xiuquan Ji
	Ivan Velitchkov Ivanov
Head of Department,	Miroslav M. Begovic

May 2018

Major Subject: Electrical Engineering

ABSTRACT

CRISPR-Cas9 has been employed as a genome editing tool in a wide range of cells of different organisms. One of the biggest challenges it faces is to maintain the efficiency of the gene regulation. To address this challenge, we have designed in this study a data-driven approach based on machine learning to predict the efficiency and to discover the mechanism of CRISPR-Cas9. We have developed Bayesian Network models to model the relationships between sequence features of target DNA and the efficiency of CRISPR-Cas9 system. We first replicated results of 2 studies and explained why naive Bayes works better as a generative model than logistic regression. Then we solved the false conditional independence of the nucleotides assumption by changing the dummy encoding to k-mer encoding. We also adopted Bayesian network structure learning and inference to assess the prediction power of the model. We eventually used D-separation analysis to study the mechanism of the CRISRR/Cas9. We combined the latest CRISPR/Cas9 structure with our D-separation analysis results and we found that the location of the active site of Cas9 and the location of scissile bonds is consistent with our D-separation findings.

DEDICATION

To my advisor, Dr. Shen, for guiding and supporting me over 2 years. To my current husband Caleb T Plant for supporting me so much and cooking for me when I was pulling all nighters. I would also like to thank my fellow researcher friends in Dr. Shen's group who gave me valuable advice. And finally, to my dog Ambrosia who kept licking me and making me feel loved.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professor Yang Shen, Professor Jean-Francois Chamberland of the Department of Electrical and Computer Engineering, Professor Jim Xiuquan Ji of the Department of Electrical and Computer Engineering, and Professor Ivan Velitchkov Ivanov of the Departments of Veterinary Physiology and Pharmacology, Electrical and Computer Engineering, and Statistics, and of the Masters programs of Biotechnology and Toxicology.

All work conducted for the thesis was completed by the student independently.

Funding Sources

There are no outside funding contributions to acknowledge related to the research and compilation of this document.

NOMENCLATURE

CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
Cas	CRISPR-associated
crRNA	CRISPR-related RNA
Cas9	CRISPR-associated protein 9
spCas9	Streptococcus pyogenes CRISPR-associated protein 9
RNA	Ribonucleic Acid
tracrRNA	Trans-activating crRNA
sgRNA	Single-guide RNA
PAM	Protospacer adjacent motif
LASSO	Least Absolute Shrinkage and Selection Operator
nt	nucleotide
A	Adenine
C	Cutosine
G	Guanine
T	Thymine
DNA	Deoxyribonucleic acid
SVM	Support vector machine
NB	Naive Bayes
bp	Base Pair
BD	Bayesian Dirichlet
BDe	Bayesian Dirichlet likelihood equivalence
BDeu	Bayesian Dirichlet equivalent uniform

DAG	Directed acyclic graph
SL	Structured learning
MRF	Markov random field
MRF	Conditional random field
AUC	Area under curve
I-map	Independency-map
P-map	Perfect-map
IS	Inter species
WL	Within library
PGM	Probabilistic graphic model

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iii

CONTRIBUTORS AND FUNDING SOURCES iv

NOMENCLATURE v

TABLE OF CONTENTS vii

LIST OF FIGURES ix

LIST OF TABLES xii

1. INTRODUCTION 1
   - 1.1 CRISPR-Cas9 for Genome Editing 1
   - 1.2 Challenges to CRISPR-Cas9 Efficiency and Previous Studies 3

2. METHODS 5
   - 2.1 Data 5
   - 2.2 Replication of Previous Studies 5
     - 2.2.1 Replication of Han's Elastic-net Regression Model 5
     - 2.2.2 Replication of Doench's Logistic Regression Classifier Model 6
   - 2.3 Bayesian Network Classifier 7
     - 2.3.1 Naive Bayes and the Comparison Between Naive Bayes and Logistic Regression 7
   - 2.4 Structure Learning for Bayesian Network 8
     - 2.4.1 Constrained and Non-Constrained Score Based Structure Learning 9
     - 2.4.2 Repeated Hill Climbing Structure Learning 9
     - 2.4.3 K-mer Encoding, False Independent Restriction Removed 10
     - 2.4.4 D-separation in Bayesian network 10

3. RESULTS AND DISCUSSION 12
   - 3.1 Replication of Elastic Net Model 12
   - 3.2 Replication of Doench Study and Comparison Between Logistic Regression and Naive Bayes 14
   - 3.3 Structure Learning and Repeated Hill Climbing Structure Learning 16

3.4 BDeu Score 18
3.5 K-mer Encoding 19
3.6 D-separation Results 23

4. SUMMARY 33

4.1 Summary of the inference results 33
4.2 Mechanism and biology research of CRISPR/Cas9 34
4.3 CRISPR/Cas9 3D structure and biophysical model 36

REFERENCES 39

LIST OF FIGURES

FIGURE Page

1.1 A simple demonstration of CRISPR/Cas9. Reprinted from [9]. Once expressed, the Cas9 protein and the gRNA form a ribonucleoprotein complex through interactions between the gRNA scaffold and surface-exposed positively-charged grooves on Cas9. Cas9 undergoes a conformational change upon gRNA binding that shifts the molecule from an inactive, non-DNA binding conformation into an active DNA-binding conformation. Eventually Cas9 creates a double strand break on the target DNA 3-nt upstream to PAM 2

2.1 An example of Naive Bayes reprinted from [18] 7

3.1 ROC curve of replication of Han's inter species experiment. Red curve is testing ROC and green curve is training ROC.AUC for training is 0.827, testing AUC is 0.812 12

3.2 ROC curve of replication Han's result -within library. Red curve is testing ROC and green curve is training ROC. AUC for training is 0.780, testing AUC is 0.846 ... 12

3.3 Replication of Elastic-net of Han and "naive Bayes-Han" 13

3.4 Replication of Doench , "naive Bayes-Han" and "naive Bayes-Doench" 16

3.5 BDeu score of "naive Bayes-Han" , "naive Bayes-Han" and "naive Bayes-Doench" .. 19

3.6 k-mer encoding AUC results and comparison with all the other AUC results 22

3.7 k-mer encoding AUC results and comparison with all the other AUC results 22

3.8 Selected replicable features for Han study and the index for non-target DNA strand in CRISPR/Cas9. Reprinted from [8] 24

3.9 Possible interacting nodes learned from constrained structure learning results. There are in total 11 experiments for both Han and Doench study, WL and IS for Han study and 9 genes for Doench study. In each experiment, we did 100 repeated structure learning. In each experiments, We counted the D-separation with frequency more than or equal to 20 out of 100. Then we find the complimentary set of these D-separation as the possible 'interacting' couples. From sub-figure A to F, the possible 'interacting' couples included are showing up in 11,10,9,8,7,6 experiments respectively. 11 is the maximum number of experiments that a possible interacting couple can show up in 25

3.10 Possible interacting nodes learned from constrained structure learning results. There are in total 11 experiments for both Han and Doench study, WL and IS for Han study and 9 genes for Doench study. In each experiment, we did 100 repeated structure learning. In each experiments, We counted the D-separation with frequency more than or equal to 100 out of 100. Then we find the complimentary set of these D-separation as the possible 'interacting' couples. From sub-figure A to F, the possible 'interacting' couples included are showing up in 11,10,9,8,7,6 experiments respectively. 11 is the maximum number of experiments that a possible interacting couple can show up in 26

3.11 Possible interacting nodes learned from unconstrained structure learning results. There are in total 11 experiments for both Han and Doench study, WL and IS for Han study and 9 genes for Doench study. In each experiment, we did 100 repeated structure learning. In each experiments, We counted the D-separation with frequency more than or equal to 20 out of 100. Then we find the complimentary set of these D-separation as the possible 'interacting' couples. From sub-figure A to F, the possible 'interacting' couples included are showing up in 11,10,9,8,7,6 experiments respectively. 11 is the maximum number of experiments that a possible interacting couple can show up in 27

3.12 Possible interacting nodes learned from unconstrained structure learning results. There are in total 11 experiments for both Han and Doench study, WL and IS for Han study and 9 genes for Doench study. In each experiment, we did 100 repeated structure learning. In each experiments, We counted the D-separation with frequency more than or equal to 100 out of 100. Then we find the complimentary set of these D-separation as the possible 'interacting' couples. From sub-figure A to F, the possible 'interacting' couples included are showing up in 11,10,9,8,7,6 experiments respectively. 11 is the maximum number of experiments that a possible interacting couple can show up in 28

3.13 Possible insignificant nodes learned from unconstrained structure learning results. There are in total 11 experiments for both Han and Doench study. In one experiment's 100 repeated structure learning results, if we found the D-separation between node X and label node, whose frequency is more than or equal to 100 out of 100, then node X will be in the output of that specific experiment. If node X is in the output of all 11 experiments, then in this histogram, the height of the bar node X will be 11. The higher the bar of node X is, more possible that node X is very insignificant to the efficiency of the CRISPR/Cas9 30

3.14 Possible insignificant nodes learned from unconstrained structure learning results. There are in total 11 experiments for both Han and Doench study. In one experiment's 100 repeated structure learning results, if we found the D-separation between node X and label node, whose frequency is more than or equal to 20 out of 100, then node X will be in the output of that specific experiment. If node X is in the output of all 11 experiments, then in this histogram, the height of the bar node X will be 11. The higher the bar of node X is, more possible that node X is very insignificant to the efficiency of the CRISPR/Cas9 31

4.1 Nucleotide type percentage change between efficient and inefficient sequences for Han data in position +1 and position N(N for NGG) 35

4.2 Nucleotide type percentage change between efficient and inefficient sequences for Doench data in position +1 and position N(N for NGG) 36

4.3 The Cas9 protein is comprised of six domains: Rec I, Rec II, Bridge Helix, RuvC, HNH, and PAM Interacting [32]. Domains are shown in schematic, crystal, and map form. Reprinted from [33] 37

LIST OF TABLES

TABLE Page

3.1 AUC of Han study and AUC of naive Bayes classifier on Han data. Replication: replication test AUC results of Han paper: elastic-net model. Naive Bayes: AUC results of naive Bayes model with 28 features 13

3.2 AUC of Bayesian network inference on Doench data 14

3.3 AUC of ROC of the model prediction on Doench testing data. Replication: replication AUC results of Doench paper. Naive Bayes- Doench: AUC results of naive Bayes model with 71 features. Naive Bayes -Han: AUC results of naive Bayes model with 28 features 14

3.4 Training and Testing AUC for Replication of Doench study 15

3.5 AUC of Bayesian network inference on Han data 17

3.6 Replication Bdeu score - of Han study and AUC of naive Bayes classifier on Han data 18

3.7 Maximum BDeu score of Doench data and AUC of naive Bayes classifier on Han data - repeated 100 times 19

3.8 Inference results of k-mer encoding of HAN data and comparison with all previous AUC results 21

3.9 Inference results of k-mer encoding of Doench data and comparison with all previous AUC results 21

1. INTRODUCTION

1.1 CRISPR-Cas9 for Genome Editing

As an analogy to text editing that is to cut, add and change texts, genome editing is a term for introducing intentional changes to the DNA sequences. DNA sequences, as central dogma states, is the source of the answers to all organisms with DNA sequences. Genome and DNA sequences may and will provide insights of how human body works and will play a vital role in health-driven research fields. The ability of making and manipulating DNA has enabled advances in biology. But introducing site-specific modifications into genome had remained elusive until the new frontier of genome editing technologies, CRISPR-Cas9 technology was discovered [1] .

The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) /Cas system was first discovered as an adaptive immune system against invading viruses in many bacteria and most archaea [2]. It was discovered that bacterias and archaea would use this system to 'memorize' the DNA sequence that virus injected into thus in the future, when the same type of virus attacks again, the bacteria or archaea would recognize it to defend themselves.

There are three types of CRISPR/Cas system [3]. While the Type I and III systems both need a more complex combination of Cas endonucleases and CRISPR related RNA (crRNA) to cleave the target DNA, the Type II system is thought to use Cas9 as the sole protein responsible for crRNA-guided silencing of foreign DNA. [4] The CRISPR type II system was discovered in Streptococcus thermophiles, a dairy bacterium. Scientist reconstituted the CRISPR system for genome editing purpose, so that it can be directed by short RNAs to induce precise cleavage in DNA [5]. The reconstituted CRISPR system is also known as CRISPR-Cas9 system, which is what this thesis is focusing on.

A CRISPR-Cas9 system only contains three components: Cas9 the nuclease, crRNA and an auxiliary trans-activating crRNA (tracrRNA)[6]. crRNA and tracrRNA can also be fused to generate a chimeric single guide-RNA(sgRNA) that mimics the natural crRNA-tracrRNA hybrid[4][7].

In a CRISPR system, an sgRNA contains a spacer sequence complementary to the targeted DNA sequence to guide the Cas9 proteins to genomic targets[8]. Once the DNA spacer is successfully targeted, the Cas9 protein will cleave and edit the DNA sequence. Figure 1.1 from [9] shows a simple demonstration of the entire process.

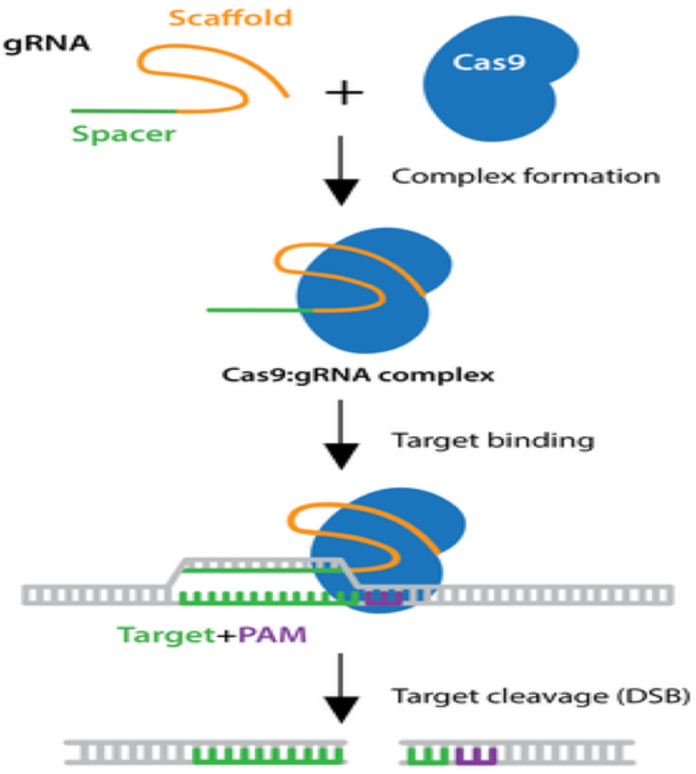


Figure 1.1: A simple demonstration of CRISPR/Cas9. Reprinted from [9]. Once expressed, the Cas9 protein and the gRNA form a ribonucleoprotein complex through interactions between the gRNA scaffold and surface-exposed positively-charged grooves on Cas9. Cas9 undergoes a conformational change upon gRNA binding that shifts the molecule from an inactive, non-DNA binding conformation into an active DNA-binding conformation. Eventually Cas9 creates a double strand break on the target DNA 3-nt upstream to PAM

In a CRISPR-Cas9 system, sgRNA or the natural crRNA and tracrRNA hybridization target a certain length of DNA sequence to allow Cas9 nuclease complex to edit the target DNA sequence. Until now, only two conditions are discovered to be necessary for sgRNA to target a DNA spacer sequence. 1) There is a unique N-nt DNA spacer sequence at the target DNA strand that can be base paired to the N-nt guide sequence within the sgRNA. N here can be 20 or other number depending on the variant that CRIPR-Cas9 system belongs to. 2) The N-nt DNA spacer sequences at non-target DNA strand sit immediately upstream to a protospacer adjacent motif (PAM)[10]. For example, PAM sequence is 5'-NGG for Streptococcus pyogenes Cas9 (spCas9). We will focus on Cas9 from S. pyogenes as it is currently the most widely used in genome engineering.

1.2 Challenges to CRISPR-Cas9 Efficiency and Previous Studies

A CRISPR-Cas9 system doesn't always work perfectly on the target DNA strand even when the mentioned two conditions are met. When the sgRNA is binded to the DNA but the sequence of DNA target spacer is not perfectly complementary to the sgRNA, we define it as a sgRNA specificity problem[11]. In a case of bad specificity, sgRNA will still be base paired although not perfectly and then degrade after the Cas9 nuclease finishes the editing/cleaving. When the sgRNA is not even paired to any DNA sequence, we define it as an sgRNA efficiency problem. In the case of low efficiency, some sgRNAs are still left intact after the experiment, which will result into a higher sgRNA abundance after the experiment than expected. In this study, we are mainly working on improving sgRNA efficiency of CRISPR-Cas9.

The cause for sgRNA low efficiency and the relationship between target DNA sequence feature and sgRNA efficiency has not been well understood yet. Despite numerous studies into this subject. One well-established belief that the conservation of proto-spacer adjacent motifs (PAMs) is a common theme for the most diverse CRISPR systems. Recently, Xu et al. found that 28 sequence features of the target DNA sequence have stronger significant impact on CRISPR knock out system by statistically analyzing the target DNA sequences with experimentally efficiencies[8], which was adopted as an valid assumption in this study. Doench et al. noticed that cytosine is preferable in PAM and some other preferences for the spacer positions adjacent to PAM for the CRISPR knockout system. They also proposed a logistic regression classifier model to discriminate the highest activity sgRNA [12]. Later, Doench et al. discovered additional features such as position-independent nucleotide counts and the location of the sgRNA target site can improve the classification result [13].

With a clear challenge statement in head, what we plan to do in this study is to use a machine learning approach combining Bayesian network parameter learning and structure learning to predict sgRNA efficiency for target DNA sequences in CRISPR-Cas9 systems. Targeting biological insights into mechanisms underlying such efficiency, we also apply structure learning and D-separation, to explain and interpret the classification results.

2. METHODS

2.1 Data

Nontarget DNA sequence strand data from 2 sources were adopted for our study. We first took the published dataset collected and included in the paper by Han Xu et al[8]. Han data can be classified into 3 categories: 40nt DNA sequence comes from ribosomal genes, non-ribosomal genes, and mESC genes which is a type of a mouse gene.

The 40-nt DNA sequence is composed by 4 parts: 3'flanking region, spacer, PAM as in NGG, and 5'flanking region. For the sequences extracted from ribosomal and non-ribosomal genes, the target spacer is 20-nt long and the 5' flanking region has 10 nucleotides. For the sequences extracted from mESC genes, the target spacer is 19-nt long and 5' flanking is 11-nt long. In all of the Han data, the 3' flanking regions are 7-nt long.

Another data we collected are from Doench et al [12]. There are 1,841 DNA sequences in Doench data from 9 genes, including human genes and mouse genes. Doench data are 30nt DNA sequences with 4 parts: 4nt 5' flanking region, 20nt spacer sequence, PAM as in NGG and 3nt 3' flanking region.

All of the DNA sequences are from the non-target DNA strand of the CRISPR-Cas9 system, which infers they are not the strand that bind with sgRNA.

2.2 Replication of Previous Studies

2.2.1 Replication of Han's Elastic-net Regression Model

Before we designed our own mathematical model to analyze the DNA sequence, the result of Han was replicated for the purpose of fair and clean comparison. In their study, an Elastic-Net model was trained to predict the efficiency of the testing sequence data. Two experiments were designed: inter species and within library. In the inter species experiment, sequences from the human ribosomal gene and human non-ribosomal gene were used to train the model, and the sequences of mESC genes from mice were used as testing data. In the within library experiment, model was trained by sequences from human ribosomal genes and then the model was tested on sequences from human non-ribosomal genes.[8]

In Han's study, the DNA sequences were represented by dummy encoding as in nucleotide A is coded as 1000, C as 0100, G as 0010 and G as 0001. Therefore, for a 40-nt DNA sequence, there will be in total 160 features. However, only 28 features were selected in Han's study to build the sequence model. These 28 features were carefully selected such that they show statistically significance and biologically reasonable correlations to the efficiency measurement of all three type of genes: ribosomal genes, non-ribosomal genes and mESC genes.[8]

Therefore, we force all the other 132 features to be 0 in our training and testing data when we were replicating the elastic net results. The labels of the data are binary as well where 1 represents inefficient and 0 represents efficient. A 3-fold cross validation was used as in the study selecting hyper-parameters for the elastic-net model.

For this replication task, we mainly used language R and R package 'glmnet' [14], which is what Han's study chose as well. For cross validation, hyper-parameter alpha is set to be between 0 and 1 with step 0.01. hyper-parameter lambda is set to be between 0.01 and 1 with step 0.01. Alpha was 0.25 and lambda is 0.01 in our study. However, according to Han's paper, their alpha is 1 which indicates their Elastic-net is Least Absolute Shrinkage and Selection Operator (LASSO) [15].

2.2.2 Replication of Doench's Logistic Regression Classifier Model

In Doench's study, 39 single nucleotide features, 30 dinucleotide features and 2 GC count features within spacer sequences were generated by L1-regularized linear support vector machine (SVM). Then a logistic regression classifier was trained to discriminate the efficient sequences. The classifier was trained on 8 genes and tested on 1 gene. Because there are in total 9, we trained 9 different logistic regression classifiers. We call this "Replication-Doench" or "Replication of Doench study".

For this replication, package 'sklearn' in Python 3 was utilized to apply logistic regression classifier. The parameters Doench chose were not specified. Eventually, we choose L1 as the regularization option and C value, inverse of regularization strength, is 1000000000[16]. This huge value of C indicates that we want L1 regularization to be applied in logistic regression to be as week as possible. This is because all the 71 features were already chosen by Doench study with L1-regularized SVM.

2.3 Bayesian Network Classifier

2.3.1 Naive Bayes and the Comparison Between Naive Bayes and Logistic Regression

NB(Naive Bayes) is a special class of Bayesian network. It applies Bayesian theorem and it indicates strong independent assumption between the features conditioned on the label [17].

To meet the assumption that NB has, the structure of it is very fixed. There will be one label node and some or one feature node. Label node is the parent node of all feature nodes. In Bayesian network, as we can see in Figure 2.1, if 2 or more nodes are not connected and they share one same parent, then they are independent from each other conditioned on the label node.

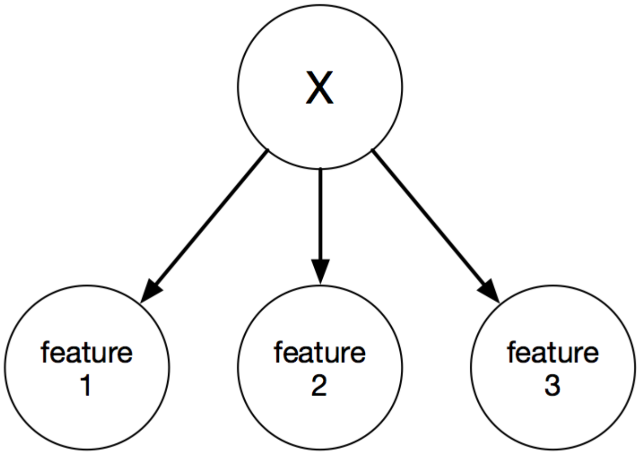


Figure 2.1: An example of Naive Bayes reprinted from [18]

To compare naive Bayes with logistic regression, we created a naive Bayes structure with all the features used in the logistic regression model, which are 39 single nucleotide features, 31 dinucleotide features and 2 GC content count features. To distinguish this particular naive Bayes model from the other naive Bayes model we will study later, we call this model as "naive Bayes - Doench". We call another naive Bayes model with only Han replicatable features, 27 features out of 30-nt sequence or in Han study 28 features out of 40-nt sequence "naive Bayes-Han". We will talk about the comparison and the results in section 3, Results and Discussion.

2.4 Structure Learning for Bayesian Network

Since one of the goals of this study is to discover the mechanism of CRISPR-Cas9 therefore we adopted Bayesian network structure learning as our tool to study the potential correlation among single nucleotides. Note that correlation does not certainly lead to causation.[19] In naive Bayes, all the feature nodes are assumed to be conditionally independent from each other. Therefore, we proposed to set naive Bayes with only single nucleotide feature nodes as the base line to compare with structure learning results.

The structure learning for Bayesian network has few basic steps. First, we learn the topology of the graph. Then we do parameter learning, in our case, because we are learning with discrete data, the parameters will be simply conditional probabilities. Then we do inference for the purpose of predicting the efficiency of DNA sequences in our study. Inference means that given some nodes, we will predict the value of some other nodes. In our study, we mainly adopted two algorithms, loopy belief propagation[20] and variable elimination[21],[22]. Loopy belief propagation is used in the cases when the undirected version of the Bayesian network is loopy, i.e. has loops. Variable elimination is used in non-loopy cases.

For Han data, the naive Bayes model has 28 feature nodes and 1 label node because we adopted the assumption that 28 single nucleotide features chosen by Han et al. are more important than other features. For Doench data, because one of the nucleotide feature chosen by Han is not included in the Doench 30nt length sequences, there are 27 feature nodes and 1 label nodes for Naive Bayes. After parameter estimation and testing, naive Bayes classifier gave a similar AUC results compared to both the replication results of Han and Doench. All the results are listed in Table 1 for comparison.

The reason of us only including 28 features into our model at this part of the study is because

Han study did not only choose 28 using a mathematical model but also biological knowledge. These 28 features are believed to have more significant impact on CRISPR-Cas9 efficiency not just in one human gene but both ribosomal and non-ribosomal genes and mouse genes[8]. Therefore, we call them replicable features.

2.4.1 Constrained and Non-Constrained Score Based Structure Learning

For the purpose of discovering relationship among feature nodes, Bayesian network structure learning was applied to the data.

Structure learning for Bayesian network is a very computationally costly problem. In this study, we used hill climbing algorithm [23] through Python package ”pgmpy” [24] for a score based structure learning where the algorithm tries to find a local maximum of the score function. The score function we used for hill Climbing was BDeu score function[25].

For constrained structure learning in BN, we set the starting structure of hill climbing structure learning to be a naive Bayes structure and we restricted the modified hill climbing algorithm so that it can not delete the naive Bayes edges. This will definitely shrink the search space of the algorithm. More specifically, we made sure that during the search, none of the 28 naive Bayes edges can be deleted, but it still can be flipped, i.e. change the direction of the edge. In addition of that, to force a wider exploration, we did not allow the algorithm to repeat 50 of the most recent actions. We name this method "constrained structure learning".

For the non-constrained structure learning method is that we allow deleting 28 Naive Bayes edges. Everything else is identical to the constrained structure learning.

2.4.2 Repeated Hill Climbing Structure Learning

Because Hill Climbing algorithm is only meant to find a local maximum point. The strategy here is to repeat the process for 100 times and find a relatively good local maxima of the Bayesian Dirichlet equivalent uniform (BDeu) score. BDeu score is one of the most used score function to assess how well the Bayesian network structure explains the dataset.

We have two purposes here. One is to testify if the structure learning results, both AUC and

BDeu score, do improve compare to naive Bayes.

The second purpose here is to discover and testify biological assumptions, which we assume that if we find some similar patterns in all of the learned high score structures, then these patterns might lead to a insightful mechanism discovery.

2.4.3 K-mer Encoding, False Independent Restriction Removed

In the previous sections, all of the nucleotides are carried out by dummy encoding, which means that if a nucleotide is A in a sequence then it is encoded as 1000, C as 0100, G as 0010, T as 0001. In this way, one nucleotide is encoded by 4 binaries. Therefore, in PGM, one nucleotide is represented by 4 nodes.

The problem with dummy encoding is that, when we build a naive Bayes model, all of the nodes are assumed to be independent from each other given label node. However, in truth it is not. It's not hard to see that for a fixed position in the DNA sequence, the nucleotides ACGT are mutually exclusive, while it is being A, it can not be any other types of nucleotides.

Therefore, we decided to change to a different type of encoding called k-mer encoding. First of all, we want to compare our previous results with new results. Therefore, we k-mer encoded the Han study and Doench study with the same chosen replicatable features. More specifically, in the previous dummy encoding experiments, 28 features are chosen from 19-nt position in the 40-nt DNA sequence. For k-mer encoding, if one position has m features that are chosen, that we are using m+1 digits to represent that position.

It is also more intuitive to analyze the structure learning results when the sequences are encoded as K-mer since now one node in the Bayesian network are one nucleotide position in DNA sequences.

2.4.4 D-separation in Bayesian network

D-separation is a criterion to determine whether node set **A** is independent of node set **B** given node set **C**. Here, we test three types of D-separation.

The first type is that if we find that node A is independent of node B given label node. The interpretation of this is that A and B are likely not interacting or affecting each other in process of CRISPR/Cas9. Even if it does, it doesn't affect the efficiency of CRISPR/Cas9.

The algorithm for finding D-separation in Bayesian network is first-breadth algorithm. Because we noticed that in repeated structure learning with the same data, all of the 100 BDeu scores are similar. We applied first-breadth algorithm to all of the 100 structures. Then we found the summation of all of the D-separation that are found in all of the 100 structures. If the frequency of some D-separation is 100, then it is the intersection among all 100 learned structures. If the frequency is 80, then it is the intersection among 80 out of 100 structures. This number will be a adjustable parameter. Eventually, we found the complementary set of these D-separation, which will be all of the possible non-conditionally independent nucleotide couples, we will call them possible interacting couples in the rest of this thesis.

The second type of D-separation analysis is that one feature node X is d-separated from label given all of the other feature nodes. The interpretation of this type of D-separation is that if all of the rest of nodes are known, then knowing node X will not further help us determine or predict the efficiency of the DNA sequence. Since our constrained structure learning did not allow any of the naive Bayes edges to be deleted, all of the feature nodes would be directly connected to the label node, thus none of the feature nodes will be conditional independent of label node. Therefore, we only conducted this second type d-separation analysis on the unconstrained results.

3. RESULTS AND DISCUSSION

3.1 Replication of Elastic Net Model

The ROC curve of the replication of Han study , i.e. elastic-net model prediction results are shown in Figure 3.1 and Figure 3.2. The AUC results of the replication are shown in Table 3.1.

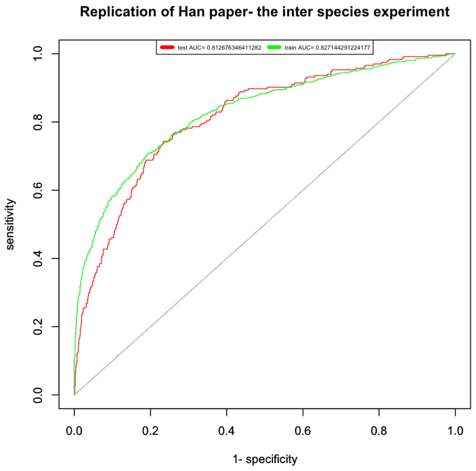


Figure 3.1: ROC curve of replication of Han's inter species experiment. Red curve is testing ROC and green curve is training ROC.AUC for training is 0.827, testing AUC is 0.812

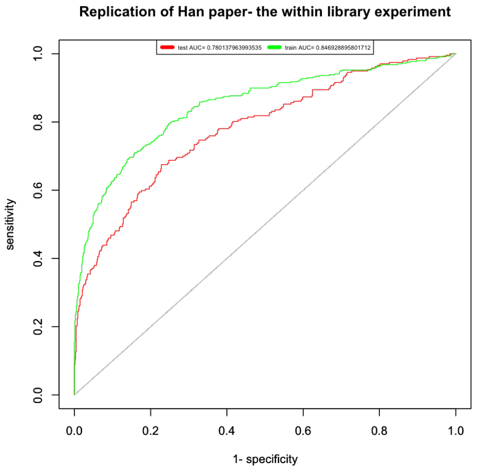


Figure 3.2: ROC curve of replication Han's result -within library. Red curve is testing ROC and green curve is training ROC. AUC for training is 0.780, testing AUC is 0.846

In Han's study, they claimed that their AUC for IS (Inter Species) is 0.757 and WL (Within Library) is 0.778. We more or less replicated their results as our IS is 0.813 and WL is 0.780.

Observing the ROC curve in Figure 3.1 and Figure 3.2 we can see that there is over-fitting in within library, which is a small surprise to me. Intuitively, we would think that if we are training and testing on the same species, human genes in this case, it would be less likely for over-fitting to happen than the inter species experiments.

This leads to a small conclusion and hyphothesis: efficiency of CRISPR/Cas9 can work more

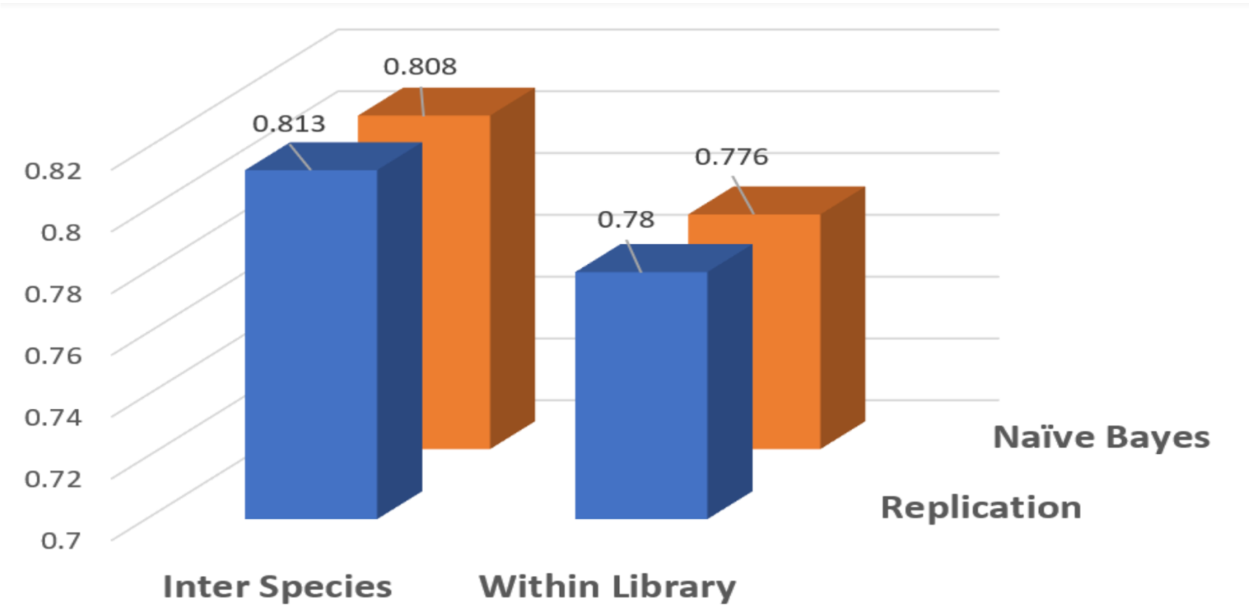


Figure 3.3: Replication of Elastic-net of Han and "naive Bayes-Han"

differently among different genes in same species than different species.

The test AUC of the Han study replication is listed in Table 3.1 in the row Replication along with some other results that we will discuss in later parts of this thesis.

Model	**Inter species**	**Within library**
Replication	0.813	0.780
Naive Bayes	0.808	0.776

Table 3.1: AUC of Han study and AUC of naive Bayes classifier on Han data. Replication: replication test AUC results of Han paper: elastic-net model. Naive Bayes: AUC results of naive Bayes model with 28 features

To have a better demonstration of the AUC of these naive Bayes inference results, we created a figure to show comparison between NB and logistic regression classification in Figure 3.3.

From the Figure 3.3, we can see that the results of naive Bayes does give very close very similar results as elastic net. This indicates that elastic net is assuming a very strong independence assumption as well. We also force some independence assumption by dummy encoded the features in this part of the study.

3.2 Replication of Doench Study and Comparison Between Logistic Regression and Naive Bayes

In Table 3.2, we listed the replication of the Doench study where they used L1 SVM for the regularization and logistic regression to train the data. We also listed results of naive Bayes inference AUC in Table 3.3, which we have mentioned in the previous section that for Naive Bayes - Doench, we chose 39 single nucleotide features selected by Doench study.And for Naive Bayes -Han, we chose 27 single nucleotide features selected by Han study.

Model	**CD13**	**CD15**	**CD28**	**CD33**	**CD43**	**CD45**	**CD5**	**H2-K**	**Thy1**
Naive Bayes-Han	0.740	0.644	0.752	0.633	0.809	0.724	0.733	0.670	0.757
one time constrained SL	0.752	0.649	0.772	0.670	0.813	0.719	0.724	0.690	0.713
one time Unconstrained SL	0.755	0.670	0.759	0.667	0.804	0.742	0.703	0.684	0.740
repeated constrained SL	0.752	0.660	0.787	0.685	0.817	0.760	0.726	0.692	0.730
repeated unconstrained SL	0.762	0.666	0.764	0.659	0.787	0.711	0.701	0.683	0.725

Table 3.2: AUC of Bayesian network inference on Doench data

Model	**CD13**	**CD15**	**CD28**	**CD33**	**CD43**	**CD45**	**CD5**	**H2-K**	**Thy1**
Replication	0.683	0.645	0.703	0.643	0.688	0.686	0.728	0.639	0.742
Naive Bayes - Doench	0.753	0.705	0.745	0.663	0.828	0.753	0.797	0.709	0.775
Naive Bayes - Han	0.740	0.644	0.752	0.633	0.809	0.724	0.733	0.670	0.757

Table 3.3: AUC of ROC of the model prediction on Doench testing data. Replication: replication AUC results of Doench paper. Naive Bayes- Doench: AUC results of naive Bayes model with 71 features. Naive Bayes -Han: AUC results of naive Bayes model with 28 features

The replication of Doench study is very close to the AUC results they published. They provide little information about their model's parameter in their paper. They only showed their results in a histogram without the actual numbers. But from their histogram, their results looks better than my AUC replication.

The most obvious finding by comparing replication of logistic regression and "naive Bayes-Doench" is that naive Bayes is doing much better than logistic regression, Even if "naive Bayes-Doench" used the exact same set of features as the logistic regression model.

The explanation for that is that because naive Bayes classifier is a typical generative model while logistic regression classifier is a discriminative classifier. For a generative model, Naive Bayes estimates the joint distribution $p(y, x)$, as in estimating the parameters for $p(y)$ and $p(x|y)$. A discriminative model, logistic regression estimates $p(y|x)$. Dr. Ng's paper talked about that the logistic regression model's estimation tends to be better than naive Bayes when the training data size is large enough. Our total dataset number is 1841. And generative model reaches asymptotic faster than a discriminative model when the training set is fewer[26].

In addition, naive Bayes has the assumption that all feature nodes are conditional independent. That gives naive Bayes a relatively small model complexity, lower variance, higher bias. This is because NB ignores correlation among the features, which induces bias and hence reduces variance. And we can also say that BN is such a simple model that it learns the parameters, which is conditional probabilities, by just calculating them. The results of naive Bayes performs better might means that data follows bias more. On the other hand, the variance of logistic regression might be too high for our dataset, which could lead to some over-fitting problem. Table 3.4 showed how the over-fitting looks like when comparing training AUC and testing AUC, where we can see that almost all training AUC are all better than testing AUC.

Model	**CD13**	**CD15**	**CD28**	**CD33**	**CD43**	**CD45**	**CD5**	**H2-K**	**Thy1**
Replication testing	0.683	0.645	0.703	0.643	0.688	0.686	0.728	0.639	0.742
Replication training	0.713	0.719	0.7113	0.718	0.709	0.710	0.711	0.715	0.709

Table 3.4: Training and Testing AUC for Replication of Doench study

To have a better demonstration of the AUC of these naive Bayes inference results, we also created a figure to show comparison between NB and logistic regression classification in Figure3.4.

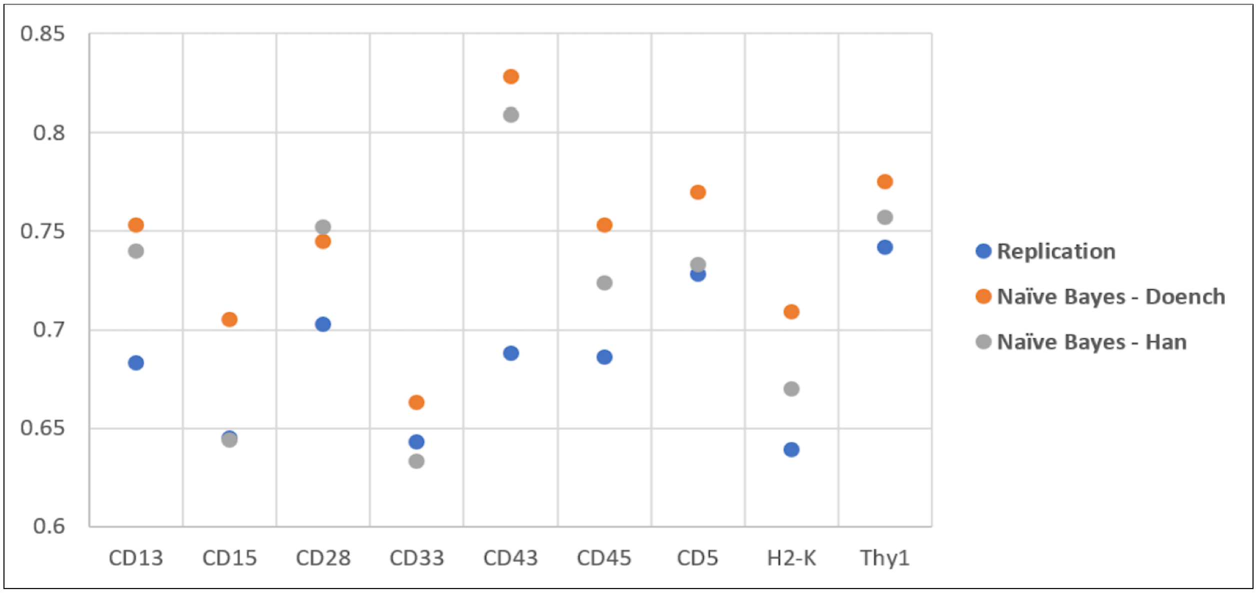


Figure 3.4: Replication of Doench , "naive Bayes-Han" and "naive Bayes-Doench"

From observing the visualization, we can have two small conclusions: 1. in Figure 3.4 we can see that the AUC of "naive Bayes-Doench" has the highest AUC in 8/9 cases. This tells that more features will give a better inference results to tell if the sequence is efficient or not. 2.

3.3 Structure Learning and Repeated Hill Climbing Structure Learning

So as we have discussed in the section 2, we first performed one-time SL(structure learning) with a BDeu score based hill climbing structure learning algorithm.

And then we performed repeated structure learning for each type of experiments the results are shown in Figure 3.5 and Figure 3.2. We have a few purposes to do this step. First, because hill-climb structure learning can only find a local optima, repeated structure learning might bring us a structure that fit the data better, a better local optima which is closer to global optima. Second, with only data and no prior knowledge or prior known causality, there are no unique structure. There can be multiple structures. If we want to analyze and explain the structure we learned, having multiple learning results is a rational thing to do.

First we can see that repeated SL does not really give a big increase in the inference AUC in both Han case and Doench case. This is due to that we are not looking for a local optimal AUC when doing SL. Our SL algorithm is score based.

Then we observed the comparison between SL and "naive Bayes-Han" where SL and "naive Bayes-Han" used the identical set of features.

For Han data, repeated constrained SL and repeated unconstrained SL did not show obvious improvement in AUC when comparing to "naive Bayes-Han". These testing AUC results are listed in Table 3.5.

For Doench data, 7 out of 9 experiments of repeated constrained SL's AUC were improved compared to "naive Bayes-Han" results by a small percentage: under 5%. 5 out of 9 experiments of repeated unconstrained SL's AUC were improved compared to "naive Bayes-Han" results by small percentage as well: under 3%. Their testing AUC results are listed in Table 3.2.

The AUC results of unconstrained SL for both Han data and Doench data is worse than constrained SL results. This indicates how important the edges between the label nodes and all replicable features nodes to the inference on the label nodes.

The reason that all these increase on AUC of ROC is not too significant could be that when we do the score based structure learning, we are searching for the structure with the highest Bdeu score, not the highest inference prediction AUC. Higher AUC of the inference on label node only means that the inference or prediction of the label node is better. But Bdeu is to assess the whole structure of the PGM, all of the nodes, not just the relationship between other feature nodes and label node. This brings us to the next question. Is the BDeu score improved?

Model	**Inter species**	**Within library**
Naive Bayes-Han	0.808	0.776
One time constrained SL	0.824	0.769
One time unconstrained SL	0.796	0.752
repeated constrained SL	0.820	0.769
repeated Unconstrained SL	0.816	0.764

Table 3.5: AUC of Bayesian network inference on Han data

3.4 BDeu Score

The BDeu score aims at maximizing the posterior probability of the DAG given data, while assuming a uniform prior over possible DAGs[27].

Because BDeu score is basically some form of log of posterior probability, they are negative.

We know that for the same set of data, the higher the Bdeu score is, the better the Bayesian network model models the joint probability distribution depending on the given data. Arguably, the most commonly used score function when performing score based BN structure learning is the BDeu, which derives from BDe and BD [27],[25],[28].

From Table 3.6, Table 3.7 and a scatter plot Figure 3.5, we observed BDeu score follows the following rule: score of unconstrained structure learning > constrained structure learning results > naive Bayes.

And the difference between naive Bayes and structure learning results are more than 1000 for both Doench and Han. The difference between constrained structure learning and unconstrained structure learning is under 100. This proves that structure learning results does improve how well the structure fit the data's joint distribution.

To explain why BDeu scores are improved when AUC is not changing much, AUC is the criteria of measuring how good the inference on only one label node is while BDeu score is a criteria of checking how the whole PGM structure is fitting the joint distribution of data. Therefore, even when the AUC of structure learning results are much better than naive Bayes, we will still analyze the results because it's possible that the structures catch some promising relationship among the feature nodes.

Model	**Inter species**	**Within library**
Naive Bayes	-33922	-19168
Constrained structure learning	-32072	-18127
Unconstrained structure learning	-32058	-18107

Table 3.6: Replication Bdeu score - of Han study and AUC of naive Bayes classifier on Han data

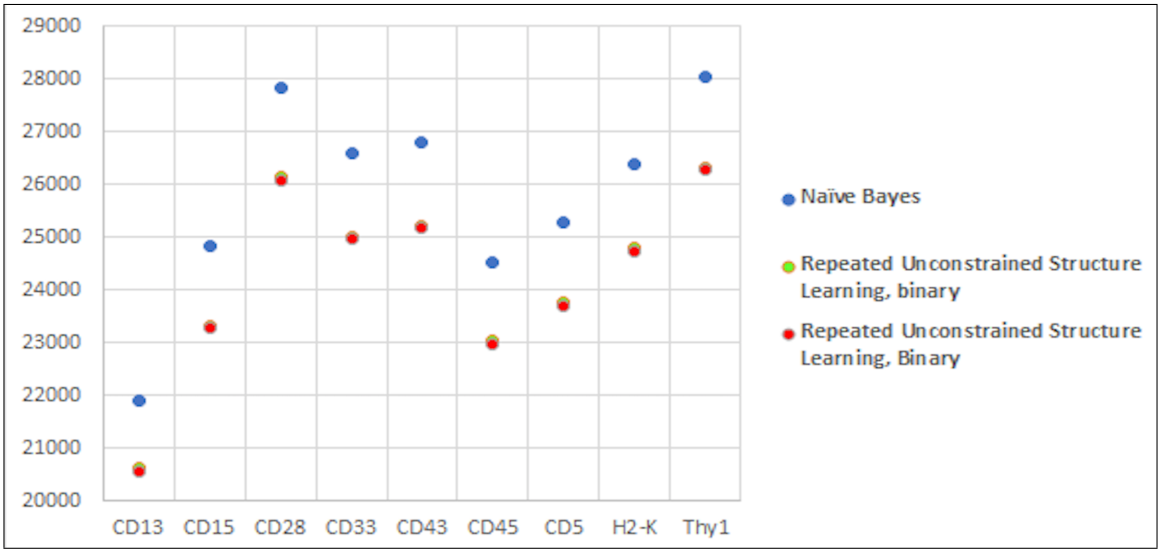


Figure 3.5: BDeu score of "naive Bayes-Han" , "naive Bayes-Han" and "naive Bayes-Doench"

Model	**CD13**	**CD15**	**CD28**	**CD33**	**CD43**	**CD45**	**CD5**	**H2-K**	**Thy1**
Naive Bayes-Han	-21897	-24840	-27836	-26597	-26815	-24526	-25276	-26401	-28033
repeated constrained structure learning,binary	-20625	-23326	-26149	-25014	-25222	-23032	-23771	-24787	-26334
repeated Unconstrained structure learning,binary	-20575	-23276	-26087	-24962	-25170	-22978	-23710	-24735	-26283

Table 3.7: Maximum BDeu score of Doench data and AUC of naive Bayes classifier on Han data - repeated 100 times

3.5 K-mer Encoding

As we mentioned in the section 2, k-mer encoding will help get rid of the false independence assumption among 4 nodes, or 4 features: ACGT in PGM. Below is the AUC of inference results of K-mer encoding sequences.

For Han study in Table 3.8, and dummy encoding results in Table 3.5, constrained structure learning for 19 nodes Bayesian network showed about a very slight increase, about 2% in AUC compared to the naive Bayes with the same number of nodes. Constrained structure learning with all 40 nodes, i.e. all positions in the given data, showed about 2% increase as well.

What is interesting is that when we look at the BDeu score in Table 3.7 and Table 3.6, we do find that the unconstrained structure learning results have a higher Bdeu score compared to constrained structure learning. But when we look at all of the AUC scores, constrained structure learning seems to usually have a better AUC. Again, as we mentioned before, this is due to that we are calculating the AUC of inference on the label node, therefore, keeping all the edges between label nodes and feature nodes. And the score-based structure learning is only trying to find a local optimal point of the BDeu score. Unconstrained SL makes the search space a bit wider.

As we mentioned in the method section, k-mer structure learning can also help us to interpret and analyze the structures much easier. For dummy encoding, one position in the DNA sequence can only be represented by 4 nodes. It's now more intuitive to look at a PGM where one node denotes to one position in DNA sequence. To show the AUC score of the inference results of K-mer decoding, two tables, Table 3.8 and Table 3.9 is shown below.

In these two tables, in few rows we mentioned 19 positions. That is because 28 features that were chosen through filtering by Han are in 19 nucleotide positions. When we are doing k-mer encoding, the input we eventually have will be 19-digits long. When in Doench study's case, there will be 18 positions due to one of the features are not included in the DNA sequences they provided.

Model	**Inter species**	**Within library**
Replicated Elastic-net with 28 features	0.813	0.780
Naive Bayes-dummy encoding with 28 features	0.808	0.776
Naive Bayes-k-mer-40 positions	0.792	0.758
constrained structure learning,40 position	0.826	0.771
Unconstrained structure learning,40 position	0.804	0.789
Naive Bayes,k-mer-19 positions	0.809	0.773
constrained structure learning,19 positions	0.827	0.792
Unconstrained structure learning,19 positions	0.823	0.778

Table 3.8: Inference results of k-mer encoding of HAN data and comparison with all previous AUC results

Model	**CD13**	**CD15**	**CD28**	**CD33**	**CD43**	**CD45**	**CD5**	**H2-K**	**Thy1**
Replication Doench- logistic regression	0.686	0.670	0.703	0.630	0.717	0.675	0.724	0.621	0.774
Naive Bayes -dummy encoding- Doench	0.753	0.705	0.745	0.663	0.828	0.753	0.797	0.709	0.775
Naive Bayes -dummy encoding- Han	0.740	0.644	0.752	0.633	0.809	0.724	0.733	0.670	0.757
Naive Bayes-18 position,k-mer	0.742	0.641	0.758	0.671	0.811	0.723	0.737	0.688	0.747
constrained SL,18 positions k-mer	0.747	0.654	0.759	0.669	0.800	0.759	0.725	0.684	0.701
Unconstrained SL,18 positions k-mer	0.740	0.649	0.777	0.652	0.775	0.739	0.707	0.689	0.733
Naive Bayes,30 positions k-mer	0.748	0.677	0.709	0.685	0.812	0.715	0.794	0.699	0.768
Unconstrained SL,30 positions k-mer value	0.740	0.649	0.777	0.651	0.775	0.688	0.707	0.689	0.732
constrained SL,30 positions K-mer value	0.747	0.654	0.759	0.669	0.800	0.759	0.725	0.684	0.701

Table 3.9: Inference results of k-mer encoding of Doench data and comparison with all previous AUC results

Again, for easier visualization, two figures were created.

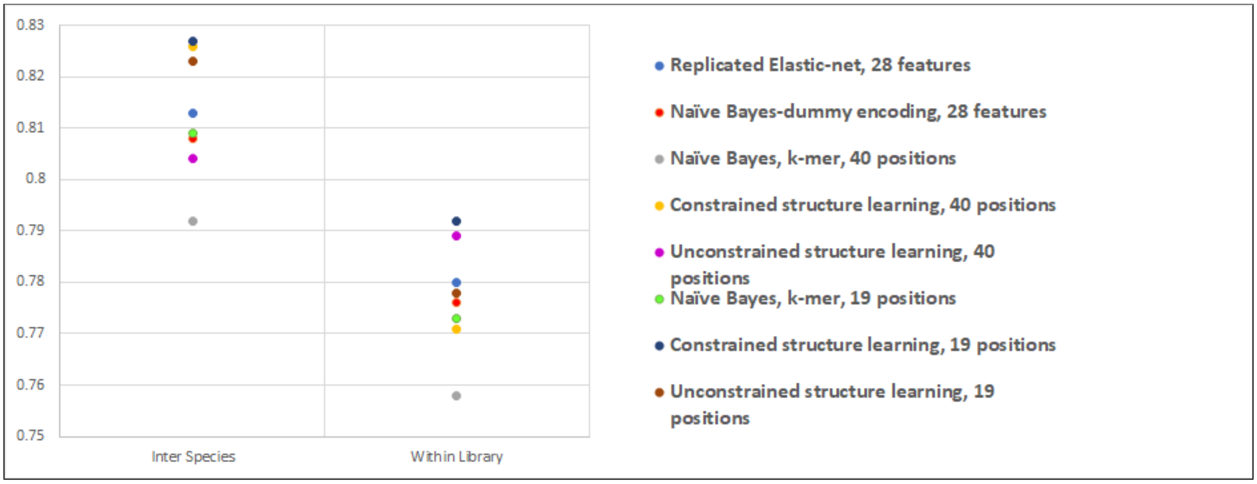


Figure 3.6: k-mer encoding AUC results and comparison with all the other AUC results

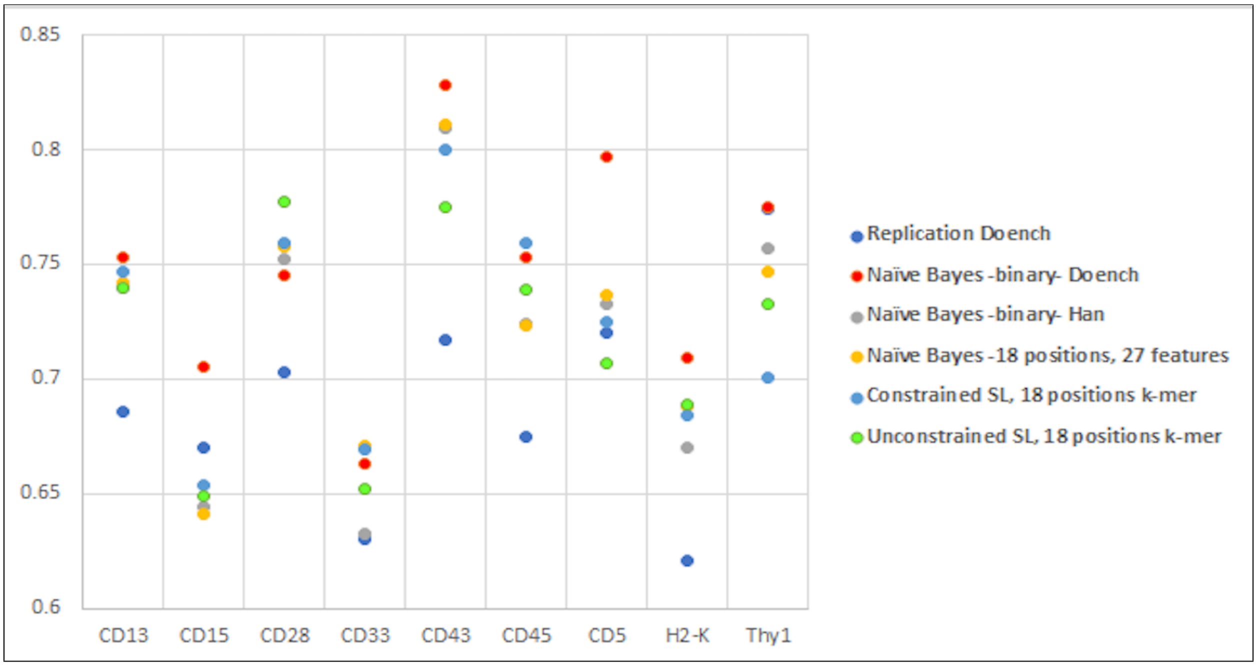


Figure 3.7: k-mer encoding AUC results and comparison with all the other AUC results

From the scatter plots Figure 3.6, we observed that "constrained SL with 19 positions and k-mer encoding" has the highest AUC in both "inter species" and "within library" experiments. "Naive

Bayes k-mer encoding on all 40 positions" have the worst AUC. The interpretation of this is that

Then we compare "constrained SL with 19 positions" and "k-mer encoding with constrained SL with 40 positions and k-mer encoding", we can see that the first one has a better AUC. The interpretation for this is that it proved that 19 positions were chosen sensibly. Adding too many single nucleotide feature might lead to over-fitting. We can also see that by comparing "Naive Bayes k-mer encoding on 19 positions" and "Naive Bayes k-mer encoding on 40 positions"

From the scatter plots Figure 3.7, we observed that replication of Doench usually gives the worst AUC , 6 out of 9 cases, which is due to the drawbacks of logistic regression model. And we also see that red dot which represents "NB with dummy encoding with all 71 features from Doench study" is still having the best AUC in most cases. This indicates that dinucleotides and GC contents are important to the efficiency of CRISPR/Cas9.

3.6 D-separation Results

Type one D-separation analyzing results are presented in Figure 3.9, and Figure 3.10. Curiously, we found out that even if we loose the restraining condition or threshold quite a bit, it didn't really change the amount of the possible interacting couples. We can see that in Figure 3.10, we only accepted those D-separations with frequency 100 which leads to a bigger complimentary set compared to Figure 3.9, where we accepted all D-separation with frequency bigger or equal to 20.

Does this mean that the shared traits of the CRISPR/Cas9 target DNA among different genes and species are really limited and specific? This can be a good news to the world of CRISPR/Cas9 study teams, this could mean that the shared similarities among different species and different genes are more rigid than we think.

Another thing that is worth mentioning is that what does "possible interacting couple" mean biologically? There are few possibilities. The first is that they are actually interacting in the 3-D structure. The second is that they are not interacting but they are very close to each other in the actual 3-D structure. The third is that they are interacting with the same third party. It's also possible that it's just a structure learning results from limited data with a limited structure learning algorithm which are only meant to find a local optima and it does not really point to anything with biological meaning.

We noticed that in all of the Type 1 D-separation results of both contrained and uncontrained structure learning results, node -1 -2 and -3 are always possibly correlated. Note that correlation does not certainly mean causation. To interpret this, we will talk more in next section.

Another interesting observation is that the most possible "interacting" couples, from sub-figure A, B to C and D of all 4 type 1 D-separation analyzing results figures, the left side of the circle is more favored than the right side, which are mostly the nucleotides in the spacer.

For quicker reference of the index of the nodes, a figure from Han's study is shown in Figure 3.8 [8] where the whole sequence is the 40-nt sequences from Ham study. In some of the positions, there are letters that represent the 28 features selected. I also add a yellow box there to indicate that from position -24 to position 3 the 30-nt sequences from Doench study.

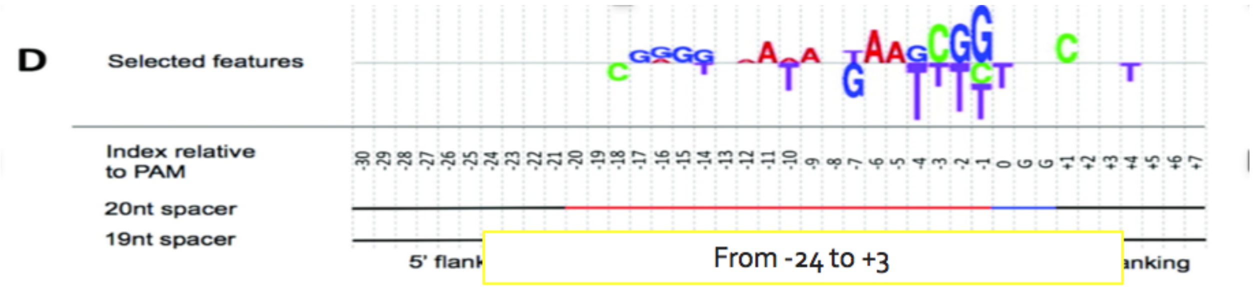

D
Selected features
Index relative to PAM
-30 -29 -28 -27 -26 -25 -24 -23 -22 -21 -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 G G +1 +2 +3 +4 +5 +6 +7
20nt spacer
19nt spacer
5' flank
From -24 to +3
nking

Figure 3.8: Selected replicable features for Han study and the index for non-target DNA strand in CRISPR/Cas9. Reprinted from [8]

.

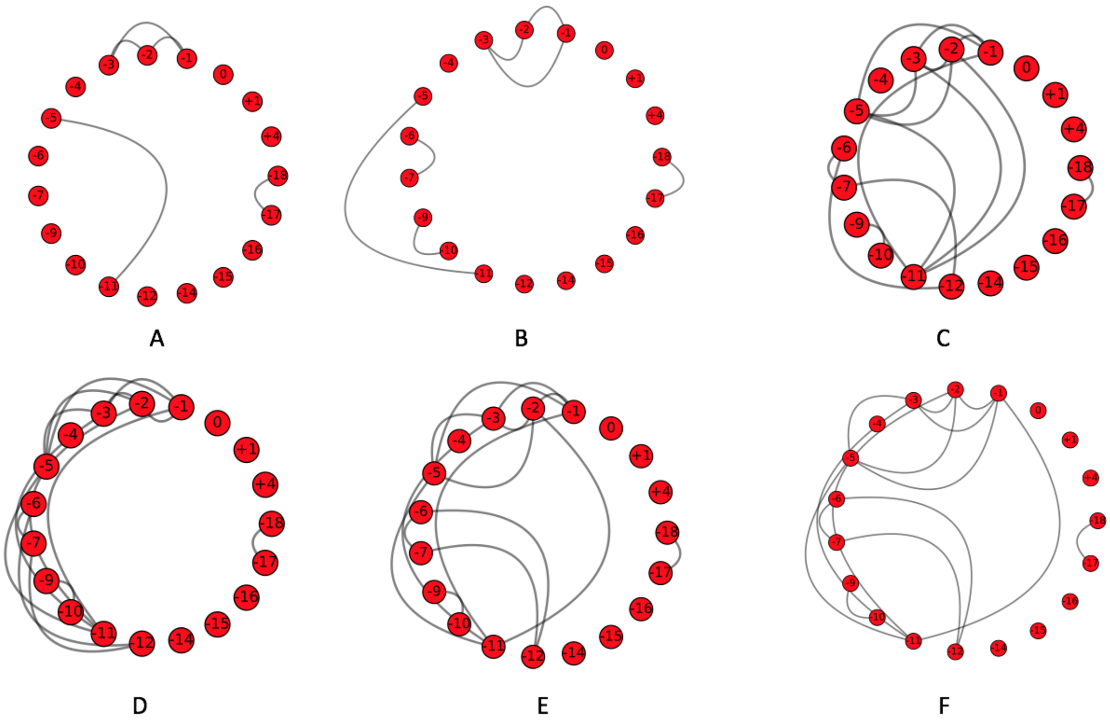


Figure 3.9: Possible interacting nodes learned from constrained structure learning results. There are in total 11 experiments for both Han and Doench study, WL and IS for Han study and 9 genes for Doench study. In each experiment, we did 100 repeated structure learning. In each experiments, We counted the D-separation with frequency more than or equal to 20 out of 100. Then we find the complimentary set of these D-separation as the possible 'interacting' couples. From sub-figure A to F, the possible 'interacting' couples included are showing up in 11,10,9,8,7,6 experiments respectively. 11 is the maximum number of experiments that a possible interacting couple can show up in

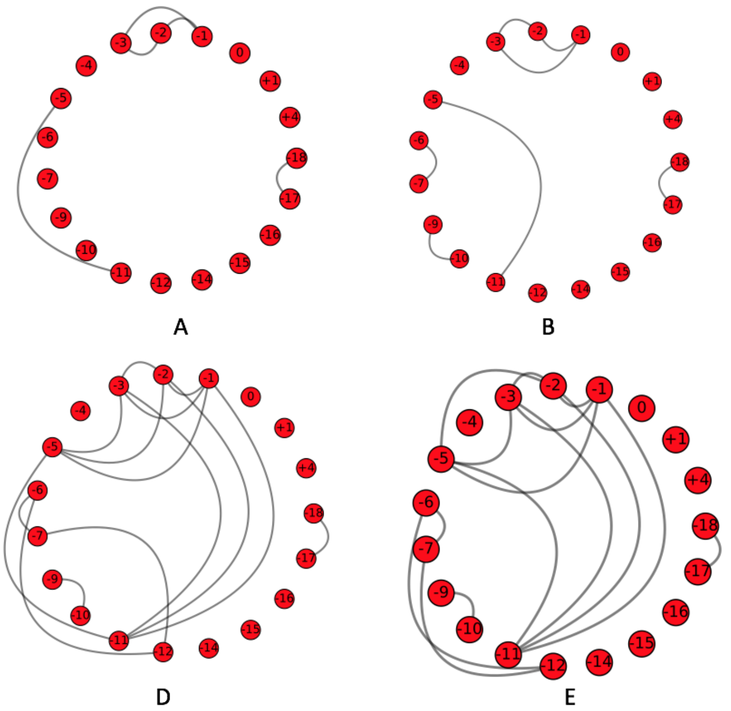


Figure 3.10: Possible interacting nodes learned from constrained structure learning results. There are in total 11 experiments for both Han and Doench study, WL and IS for Han study and 9 genes for Doench study. In each experiment, we did 100 repeated structure learning. In each experiments, We counted the D-separation with frequency more than or equal to 100 out of 100. Then we find the complimentary set of these D-separation as the possible 'interacting' couples. From sub-figure A to F, the possible 'interacting' couples included are showing up in 11,10,9,8,7,6 experiments respectively. 11 is the maximum number of experiments that a possible interacting couple can show up in

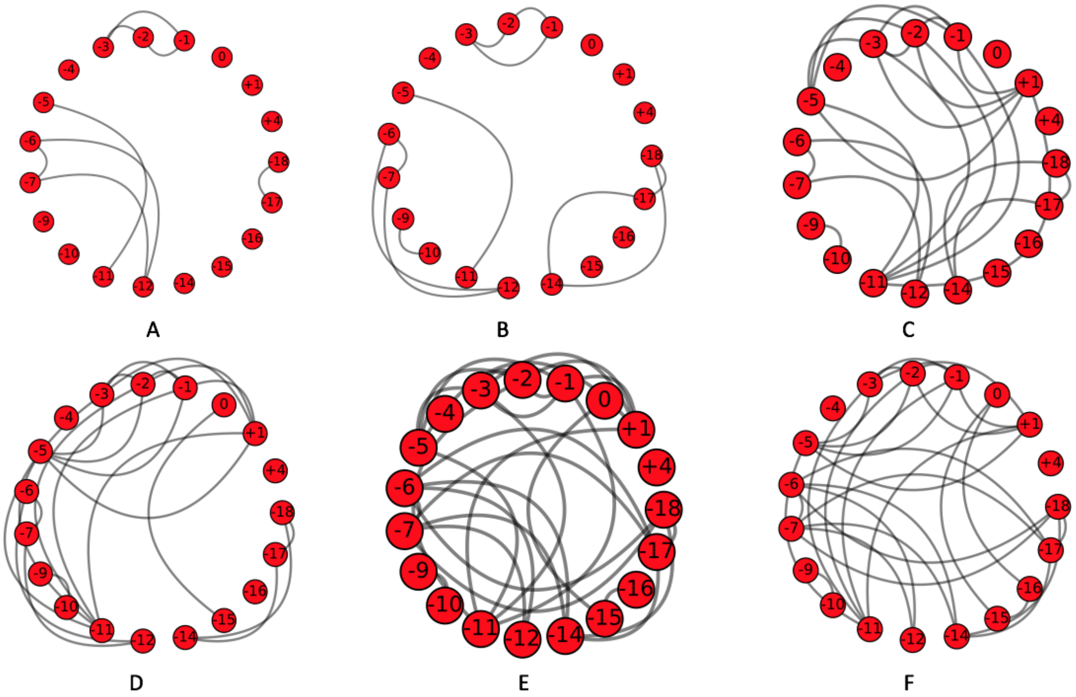


Figure 3.11: Possible interacting nodes learned from unconstrained structure learning results. There are in total 11 experiments for both Han and Doench study, WL and IS for Han study and 9 genes for Doench study. In each experiment, we did 100 repeated structure learning. In each experiments, We counted the D-separation with frequency more than or equal to 20 out of 100. Then we find the complimentary set of these D-separation as the possible 'interacting' couples. From sub-figure A to F, the possible 'interacting' couples included are showing up in 11,10,9,8,7,6 experiments respectively. 11 is the maximum number of experiments that a possible interacting couple can show up in

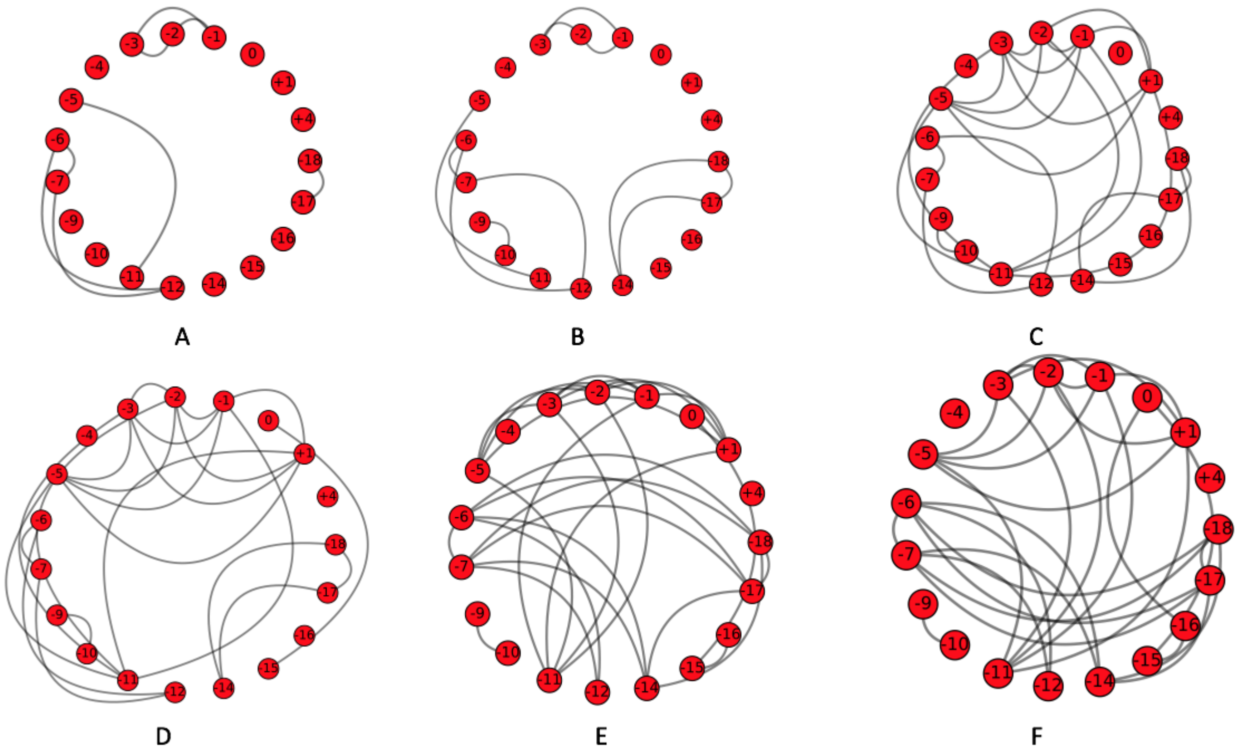


Figure 3.12: Possible interacting nodes learned from unconstrained structure learning results. There are in total 11 experiments for both Han and Doench study, WL and IS for Han study and 9 genes for Doench study. In each experiment, we did 100 repeated structure learning. In each experiments, We counted the D-separation with frequency more than or equal to 100 out of 100. Then we find the complimentary set of these D-separation as the possible 'interacting' couples. From sub-figure A to F, the possible 'interacting' couples included are showing up in 11,10,9,8,7,6 experiments respectively. 11 is the maximum number of experiments that a possible interacting couple can show up in

For Type 2 D-separation where we find position that is d-separated from label node given all the other feature nodes. Let Y denote label node in the BN. Equation 3.1 is the mathematical expression for this type of D-separation.

$$X_i \perp\!\!\!\perp Y \mid allX_j, forj \neq i \tag{3.1}$$

The interpretation of this equation is that we consider X_i to be a node possibly contributing less or nothing to the efficiency of the DNA sequence in CRISR/Cas9 system. For convenience purpose, we call these X_i insignificant nodes.

We choose the word 'possibly' in the last paragraph for a reason. Firstly, even in a good I-map BN, which means that this BN's d-separation is a subset of actual conditional-independence in the data, and when it is a good I-map(independence map), it covers most of the independence in the data, sometimes we can still have independence that can not be captured by the BN, which means it will not show up in the D-separation. This is also known as P-map(perfect map) does not always exist.

In addition to that, our data is just samples. It is limited. They are truly just small amount of samples of the non-target strand out of the whole world's possible CRISPR/Cas9 DNA targets sequences. A small number of data can possibly cause the Bayesian network structure learning to capture some false conditional independence which does not really exist in real data (infinite number of data).

But it doesn't mean that these D-separation analyzing results are meaningless. We still need to believe with thousands of data, BN structure learning and D-separation could give us some good insights.

We plot the results of type 2 d-separation in Figure 3.13 and Figure 3.14

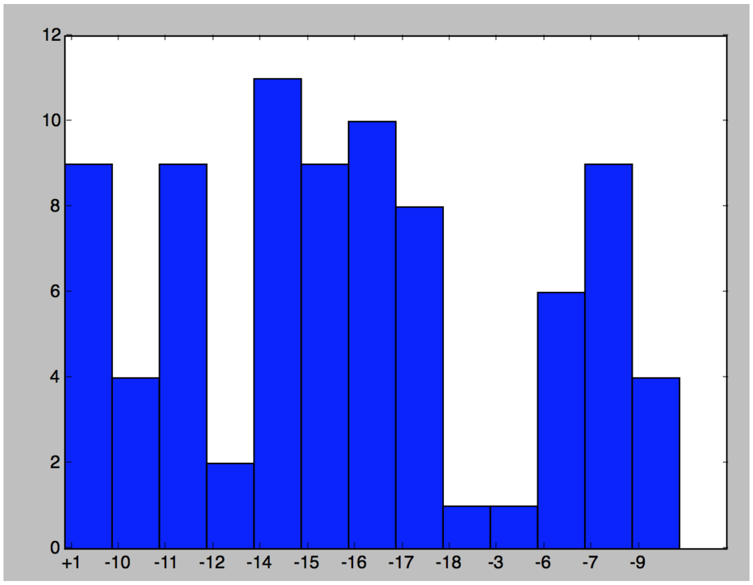


Figure 3.13: Possible insignificant nodes learned from unconstrained structure learning results. There are in total 11 experiments for both Han and Doench study. In one experiment's 100 repeated structure learning results, if we found the D-separation between node X and label node, whose frequency is more than or equal to 100 out of 100, then node X will be in the output of that specific experiment. If node X is in the output of all 11 experiments, then in this histogram, the height of the bar node X will be 11. The higher the bar of node X is, more possible that node X is very insignificant to the efficiency of the CRISPR/Cas9

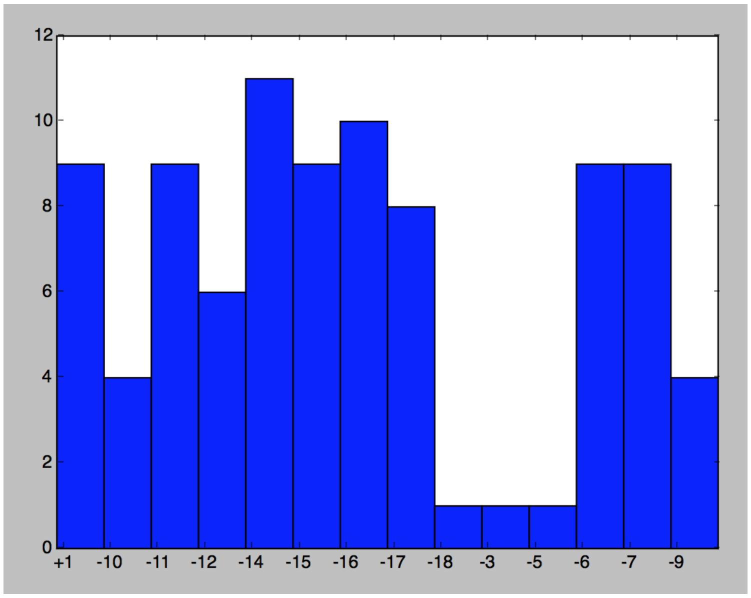


Figure 3.14: Possible insignificant nodes learned from unconstrained structure learning results. There are in total 11 experiments for both Han and Doench study. In one experiment's 100 repeated structure learning results, if we found the D-separation between node X and label node, whose frequency is more than or equal to 20 out of 100, then node X will be in the output of that specific experiment. If node X is in the output of all 11 experiments, then in this histogram, the height of the bar node X will be 11. The higher the bar of node X is, more possible that node X is very insignificant to the efficiency of the CRISPR/Cas9

To be very rigorous, we can't say that these nodes in the histogram have nothing to do with the mechanism of CRISPR/Cas9. It might still have correlations with other nodes. For an example, when the rest of the sequence is set, position X_i has to be T, but sequence can be either efficient or inefficient.

Taking a look at Figure 3.13 and Figure 3.14. In the analyzing results of type 2 d-separation results, when we set the threshold of the frequency of d-separation to be 20 (loose, more d-separation accepted), in Figure 3.14, we noticed that most of the insignificant nodes with taller bars are-14,-16,-15,-7,+1,-11. For the second case where we set the threshold of the frequency of d-separation to be 100 (tight, less d-separation accepted), we noticed that most of the insignificant nodes with taller bars are -17, -16, -15, -11.

From Figure 3.8, we can see that from index -20 to -10, these are the positions within the spacer that’s closer to the 5’ flanking side. And our type 2 D-separation shows that nucleotides closer to the 5’ end are less likely to play an important role in CRISPR/Cas9.

4. SUMMARY

4.1 Summary of the inference results

In summary of the prediction or inference results, all the current AUC results indicate that it is difficult to beat naive Bayes classifier's AUC, which is the reason why we compare a lot of the results with naive Bayes classifier's AUC. Since naive Bayes indicates really strong independence assumption, we suspect that most of the chosen nucleotides are indeed independent from each other given the efficiency of the sequence, the label node. And by applying hill climbing structure learning algorithm, because of the limitation of it only being able to find a local maximum, it is hard to have a much better results than naive Bayes. However, we did find in all cases, a constrained structure learning's inference usually improves the AUC by 2% to 4%. And a constrained structure learning's inference's AUC is usually better than unconstrained structure learning. What does that mean? Didn't we find that the Bdeu score of unconstrained is higher than constrained learning?

Intuitively, I suspect it indicates that because constrained structure learning stops the algorithm from deleting any naive Bayes edges, the inference path to the label node is more straight forward than unconstrained structure learned results. We basically assumed that in constrained structure learning, all feature nodes are dependent with the label node.

However, unconstrained structure learning is still valuable for us to conduct the type 2 D-separation analysis. With constrained learning keeping all edges between label node and features nodes, the type 2 D-separation analysis can not be achieved.

Another important finding is that based on our results, we do believe that 28 features chosen from 19 position in Han study is valuable for future CRISPR/Cas study. We also believe that GC content and dinucleotides are valuable features.

In addition, we noticed that for Doench study, naive Bayes shows clear improvement compared to logistic regression classifier. Our interpretation, as stated in a previous section, is that because naive Bayes is a typical generative model while logistic regression classifier is a discriminative classifier. It is usually harder to train a generative model but in this case, naive Bayes is a very simple generative model with not many parameters. Also because of the data size, naive Bayes

Yet the goal of this study is not just to beat the AUC of the classifier but also help reveal and understand the mechanism of CRISPR.

4.2 Mechanism and biology research of CRISPR/Cas9

Obviously, we want to answer the following ultimate question: what can we do to make CRISPR/Cas9 a more efficient and precise genome editing tool for not just simple species, but all species.

In our study, before facing that ultimate question, we would like to firstly answer some smaller questions to pave the way to a bigger picture. For example, does structure learning of Bayesian network really help us understanding CRISPR/Cas9 more? Does d-separation results supports other scientists' findings? Does d-separation results proposed new insights of the mechanism of CRISPR/Cas9?

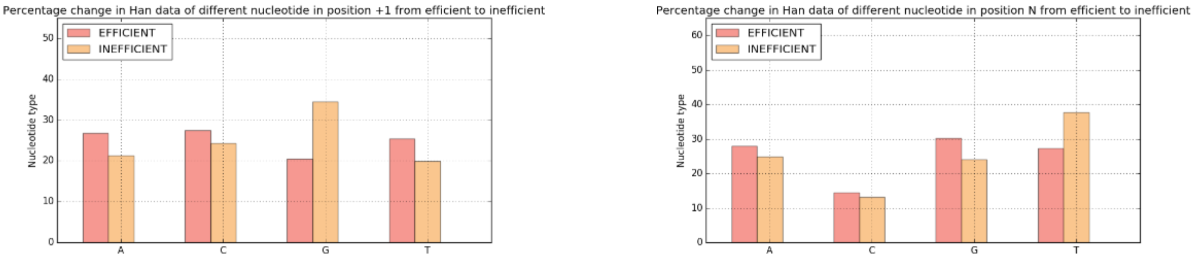
In Doench paper, they mentioned that a bias against thymine towards the 3' end of the 20-nt sgRNA target site and others has previously been explained from the perspective of sgRNA expression. They also observed a strong bias against guanine immediately 3' of the PAM [12].

Doench study also noticed a preference in the variable nucleotide of the PAM, where cytosine was favored and thymine was disfavored, which is consistent with Han's replicatable features where they select thymine as a negative feature [12].

Curiously, most of these observations, favoring of the nucleotides from Doench study are usually close to the 3' side of the sequence or the spacer. And in our study, we did notice from the results of type 2 D-separation analysis that 5' side of the spacer are less significant to the mechanism of CRISPR/Cas9. We also noticed from type 1 D-separation that -1 , -2 and -3 , which are the spacer nucleotides right upstream to the PAM are always interacting with each other in all analysis results.

Another interesting knowledge of the CRISPR/Cas9 is that between -3 and -4 are the cleavage position of spCas9 [9]. And we did notice the correlation among -1, -2 and -3 from type 1 d-separation analysis, which might be due to the cleavage site of the Sp-Cas9 nuclease. Sp-Cas9 nuclease HNH domain cleaves the phosphodiester bond (P-O) between -3 and -4[4] [29].

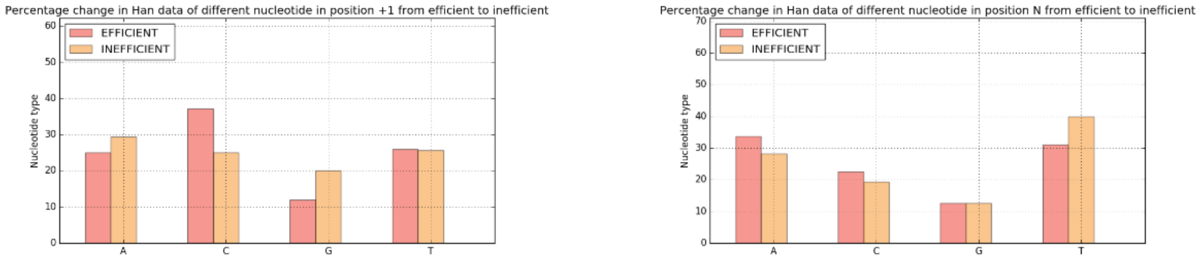
In a study from Pennsylvania State University, they found that the first nucleotide in PAM (N in NGG) did not significantly contribute to Cas9's cleavage activity, but that the fourth nucleotide, whose index in our case is +1 did significantly alter cleavage activity[30]. We couldn't find any of our results supporting this finding but a D-separation analysis might give some ideas. In fact, the analysis results of our type 2 d-separation supports the idea that changing the nucleotide choice in position +1 would not alter the activity cleavage much. This could be due to the limit amount of the sample we have or different type of data we use. To further test this, we calculated the percentage of different nucleotides position +1 of effective sequences and ineffective sequence. The calculation results are shown in Figure 4.1 and Figure 4.2. From comparing the percentage change, I don't see a huge difference between position +1 change and position N change (or position 0) from efficient to inefficient sequence.



(a) Percentage change for position +1

(b) Percentage change for position N

Figure 4.1: Nucleotide type percentage change between efficient and inefficient sequences for Han data in position +1 and position N(N for NGG)



(a) Percentage change for position +1

(b) Percentage change for position N

Figure 4.2: Nucleotide type percentage change between efficient and inefficient sequences for Doench data in position +1 and position N(N for NGG)

4.3 CRISPR/Cas9 3D structure and biophysical model

We have found a few possible long range correlation from our d-separation analysis. For example, correlation between -5 and -11, correlation between -6 and -12 and some close correlation among -1,-2 and -3. We want to see if the research on the structure of the CRISPR/Cas9 will be consistent with our findings.

After studying Cas9-mediated DNA site cleavage, researchers have divided it into a few stages: (1) Forming active crRNA (2) Cas9 binding to active crRNA (3) Resulting active complex performs a movement, it may or may not successfully bind to a DNA sequence. The success rate is determined by PAM , number of same sequence DNA sites and their binding free energy. (4) The formation of a stable Cas9:crRNA:DNA complex. For step 4, the Cas9:crRNA recognize PAM and then DNA duplex will need to be unwinded. And then RNA pairs with target DNA strand which leads to a DNA:RNA:DNA complex called R-look. (5) After cleavage, the Cas9:crRNA:DNA complex remains bound to the cleaved DNA, and is considered a no-turnover enzyme [30].

As we can see, to fully understand CRISPR/Cas9 and the mystery of its imperfect activity and off-target behavior, structures at different states will all need to be throughly studied.

A recent study of Huai showed us the structure of Cas9 in complex with sgRNA and target DNA and it reveals the overall topology of the non-target strand bound to SpCas9[31]. For easier understanding and demonstration, a figure as Figure 4.3 is included to show domains of Cas9.

Another figure is included to shown

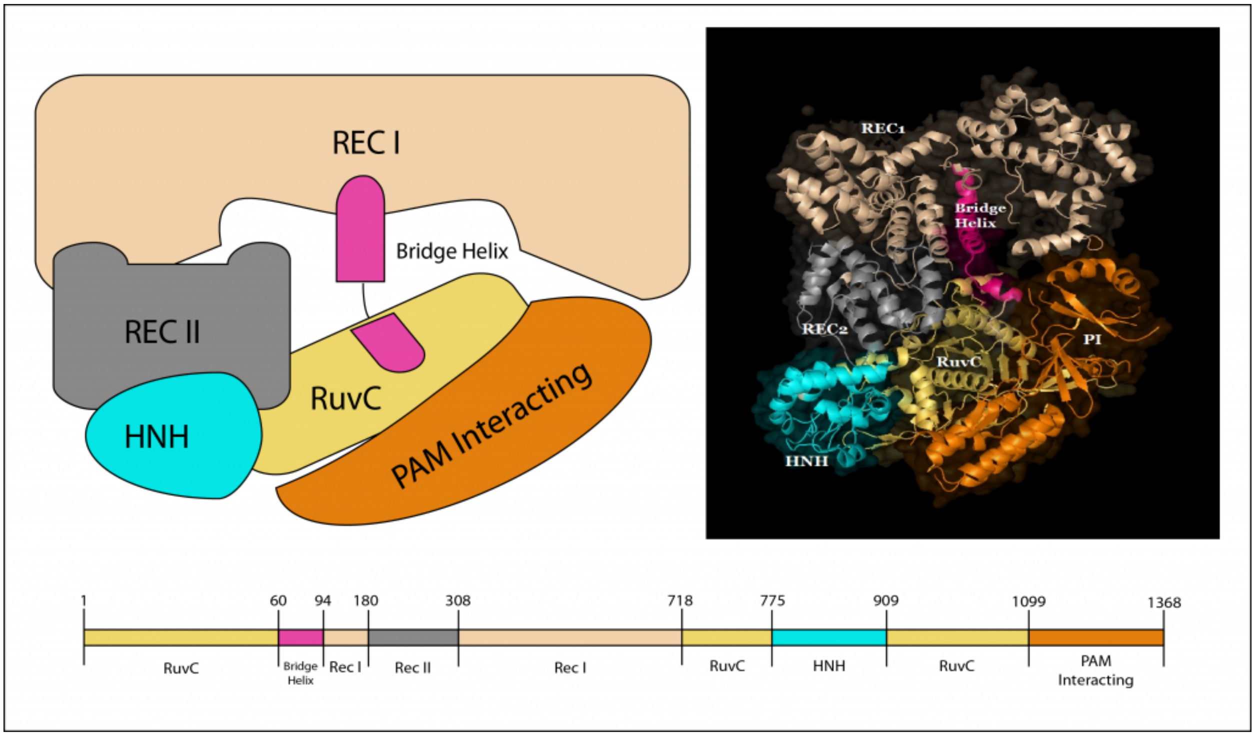


Figure 4.3: The Cas9 protein is comprised of six domains: Rec I, Rec II, Bridge Helix, RuvC, HNH, and PAM Interacting [32]. Domains are shown in schematic, crystal, and map form. Reprinted from [33]

.

According to Huai study, they revealed a distinct SpCas9 state in which the HNH active site is the closest to the scissile bond of the target DNA[31]. Scissile bond is basically the cleavage site where between position -3 and position -4 the two nuclease domains (HNH and RuvC) catalyze the splitting of the scissile bonds in two DNA strands, respectively. This could also possibly explain the' interacting' between position -3, -2 and -1. They are all close to the HNH active site.

Their study also showed base -1 to -7 of the non-target strand are enclosed in the channel formed by the HNH, RuvC and PI domains, while the rest 13 bases from -8 to -20 are bound to the surface of Spcas9. This is consistent to our findings of the type 2 d-separation analysis results to a degree. In Figure 3.13 most of the insignificant nodes are from -8 to -20. Only 4 out of 13 are not in this range. But with my limited biology knowledge, I will need further study to understand it.

Huai study also revealed an interaction site of the 16-bp PAM proximal end with SpCas9, which would be from position +1 to +16. It can not be validated by our model because our sequence only has nucleotide position from +1 to +7.

REFERENCES

[1] J. A. Doudna and E. Charpentier, ”The new frontier of genome engineering with CRISPR-Cas9” *Science*, vol. 346, no. 6213, p. 1258096, Nov. 2014.

[2] H. Deveau, J. E. Garneau, and S. Moineau, ”CRISPR/Cas system and its role in phage-bacteria interactions,” *Annu. Rev. Microbiol.*, vol. 64, pp. 475-493, 2010.

[3] K. S. Makarova et al., ”Evolution and classification of the CRISPR-Cas systems,” *Nat Rev Micro*, vol. 9, no. 6, pp. 467-477, Jun. 2011.

[4] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, ”A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity,” *Science*, vol. 337, no. 6096, pp. 816-821, Aug. 2012.

[5] L. Cong et al., ”Multiplex Genome Engineering Using CRISPR/Cas Systems,” *Science*, vol. 339, no. 6121, pp. 819-823, Feb. 2013.

[6] E. Deltcheva et al., ”CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III,” *Nature*, vol. 471, no. 7340, pp. 602-607, Mar. 2011.

[7] P. D. Hsu et al., ”DNA targeting specificity of RNA-guided Cas9 nucleases,” *Nat Biotech*, vol. 31, no. 9, pp. 827-832, Sep. 2013.

[8] H. Xu et al., ”Sequence determinants of improved CRISPR sgRNA design,” *Genome Res.*, p. gr.191452.115, Jun. 2015.

[9] ”Addgene: CRISPR Guide.” [Online]. Available: https://www.addgene.org/crispr/guide/. [Accessed: 03-Mar-2018].

[10] F. J. M. Mojica, C. Díez-Villaseñor, J. García-Martínez, and C. Almendros, ”Short motif sequences determine the targets of the prokaryotic CRISPR defence system,” *Microbiology (Reading, Engl.)*, vol. 155, no. Pt 3, pp. 733-740, Mar. 2009.

[11] X. Wu, A. J. Kriz, and P. A. Sharp, ”Target specificity of the CRISPR-Cas9 system,” *Quant Biol*, vol. 2, no. 2, pp. 59-70, Jun. 2014.

[12] J. G. Doench et al., ”Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation,” *Nat Biotechnol*, vol. 32, no. 12, pp. 1262-1267, Dec. 2014.

[13] J. G. Doench et al., ”Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9,” *Nat Biotech*, vol. 34, no. 2, pp. 184-191, Feb. 2016.

[14] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/, (2010)

[15] R. Tibshirani, ”Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society*. Series B (Methodological), vol. 58, no. 1, pp. 267-288, 1996.

[16] F. Pedregosa, and G. Varoquaux, and A. Gramfort, and V. Michel, and B. Thirion, and O. Grisel, and M. Blondel, and P. Prettenhofer, and R. Weiss, and V. Dubourg, and J. Vanderplas,

and A. Passos, and D. Cournapeau, and M. Brucher, and M. Perrot, and E. Duchesnay, ”Scikit-learn: Machine Learning in Python”, *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011

[17] ”Properties of Naive Bayes.” [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/properties-of-naive-bayes-1.html. [Accessed: 22-Mar-2018].

[18] ”PR3 Naive Bayes | Machine Learning.” Accessed March 23, 2018. http://www.cs.cornell.edu/courses/cs4780/2015fa/iframe/iframe-5/index.html.

[19] D. Koller, and N. Friedman, ”Probablistic graphic model: principle and techniques”, *The MIT Press*, 2009

[20] K. P. Murphy, Y. Weiss, and M. I. Jordan. ”Loopy belief propagation for approximate inference: An empirical study.” *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence.* Morgan Kaufmann Publishers Inc., 1999.

[21] B. D’Ambrosio, ”Inference in Bayesian Networks,” p. 16.

[22] M. Chavira and A. Darwiche, ”Compiling Bayesian Networks Using Variable Elimination,” *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, San Francisco, CA, USA, 2007, pp. 2443âĂŞ2449.

[23] I. Tsamardinos, L. Brown, and C. F. Aliferis. ”The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm.” *Machine Learning 65*, no. 1 (October 1, 2006): 31-78. https://doi.org/10.1007/s10994-006-6889-7.

[24] Pgmpy: Python Library for Probabilistic Graphical Models. Python. 2013. Reprint, pgmpy, 2018. https://github.com/pgmpy/pgmpy.

[25] W. Buntine, ”Theory Refinement on Bayesian Networks,” *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 52-60, 1991

[26] A. Y. Ng, and M. I. Jordan, ”On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes,” in Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, Cambridge, MA, USA, 2001, pp. 841-848

[27] M. Scanagatta, C. P. de Campos, and M. Zaffalon, ”Min-BDeu and Max-BDeu Scores for Learning Bayesian Networks,” *Probabilistic Graphical Models*, vol. 8754, L. C. van der Gaag and A. J. Feelders, Eds. Cham: Springer International Publishing, 2014, pp. 426-441.

[28] G. F. Cooper, and E. Herskovits, ”A Bayesian Method for the Induction of Probabilistic Networks from Data,” *Machine Learning 9*, pp. 309-347, 1992

[29] G. Gasiunas, R. Barrangou, P. Horvath, and V. Siksnys, ”Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 39, pp. E2579-2586, Sep. 2012.

[30] I. Farasat and H. M. Salis, ”A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation,” *PLOS Computational Biology*, vol. 12, no. 1, p. e1004724, Jan. 2016.

[31] C. Huai et al., ”Structural insights into DNA cleavage activation of CRISPR-Cas9 system,” *Nature Communications*, vol. 8, no. 1, p. 1375, Nov. 2017.

[32] C. Anders, O. Niewoehner, A. Duerst, and M. Jinek, ”Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease,” *Nature*, vol. 513, no. 7519, pp. 569-573, Sep. 2014.

[33] ”CRISPR Mechanism | CRISPR/Cas9.” [Online]. Available: https://sites.tufts.edu/crispr/crispr-mechanism/. [Accessed: 23-Mar-2018].